United Nations ECE/CES/2015/39

Economic and Social Council

Distr.: General
4 June 2015

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-third plenary session
Geneva, 15-17 June 2015
Item 4 of the provisional agenda
Modernisation of statistical production and services and managing for efficiency

Building a common production environment for the statistical production process

Note by Statistics Sweden

Summary

This paper gives a short overview of the progress of Statistics Sweden in building a common statistical production environment. Furthermore, the paper describes the ongoing work and elaborates on some of the lessons learned during the process of introducing a common statistical production process.

The paper is presented for discussion to the second session of the Conference of European Statisticians' seminar "Modernisation of statistical production and services and managing for efficiency".

GE.15-08985 (E)

1508985

Please recycle

I. Introduction

1. Since 2008, Statistics Sweden has moved to a process-oriented approach in statistical production. The aim is to make production more efficient and standardised, using common methods and tools wherever possible. One important aspect of this is building a common information technology (IT) platform where existing and new tools can communicate efficiently with each other, and where statisticians can choose the necessary tools for each survey. For data collection, this platform has been realised through the so-called TRITON project, which is now used by many surveys to collect and validate data through web collection and scanned paper forms. The structure set by the TRITON project is now being applied and expanded to other parts of the statistical production process (processing, analysis and dissemination). The paper describes how far Statistics Sweden has come, describes the ongoing work and elaborates on some of the lessons learned so far.

II. The current situation and development so far

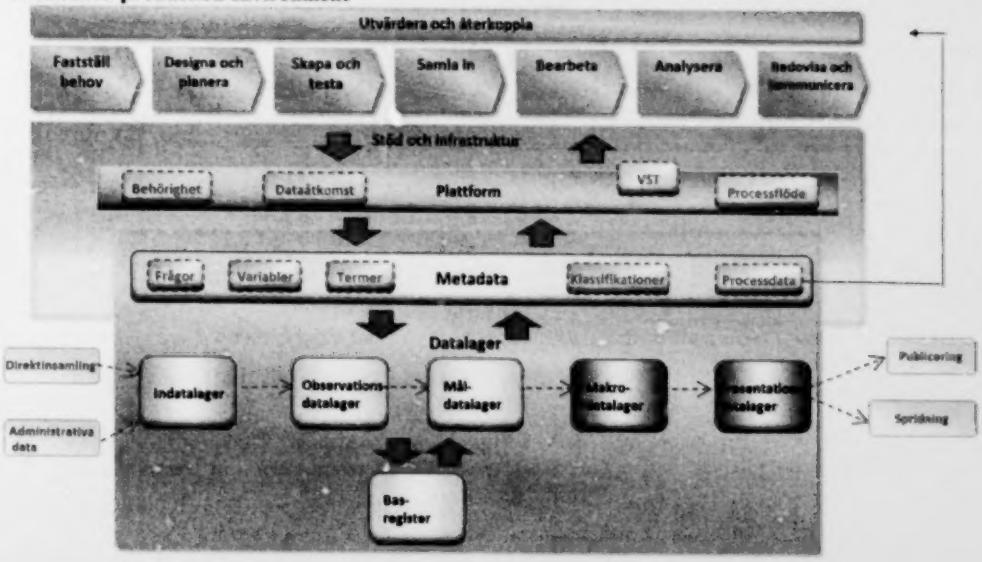
2. In 2008, Statistics Sweden introduced a process-oriented approach to the statistical production process, introducing our own version of the Generic Statistical Business Process Model (GSBPM, see Figure 1 below). The adapted GSBPM was divided into five sub-process groups, which were all given a devoted process owner (and a deputy). The role of the process owner is to lead the development and maintenance of methods, tools and routines within his or her sub-process, to give support and to provide guidelines and descriptions of the process and how to run it.

3. The descriptions and guidelines on how to run the statistical production process were put on the internal website of Statistics Sweden in a tool called the process support system (PSS). PSS can be seen as a handbook in statistical production at Statistics Sweden, and includes all agreed-upon common methods, IT tools, routines, checklists, templates etc. that the different surveys must use in their individual production processes. But the vision for PSS has from the start been to create something that is even more active, giving each survey more specific information based on the characteristics of the particular survey.

4. At the same time, there was a need for modernisation of many of the survey specific IT tools used – these being a result of the decentralised organisation we had and the need for development around the millennium shift where a personal computer (PC)-based platform was introduced and many systems needed to be replaced quickly. This resulted in many local systems. Now, around 10 years later, the drawbacks of these systems were becoming clear – dependency on specific persons, old technical solutions that were no longer in use or supported, and expensive maintenance. The drawbacks were especially notable since there were new, more common solutions being built (for example for web data collection) that had to be integrated with all the survey specific systems. So the idea was raised to combine the initiatives and start building a common production environment with the parts outlined in the conceptual picture in Figure 2 (in Swedish). With the production process at the top and a new, structured data storage function at the bottom, a number of IT tools or services were envisioned to access the data and enhance it, with metadata being used to connect the tools and the data. The IT tools are accessed through a new version of the PSS (or VST in Swedish, as can be seen in the picture).

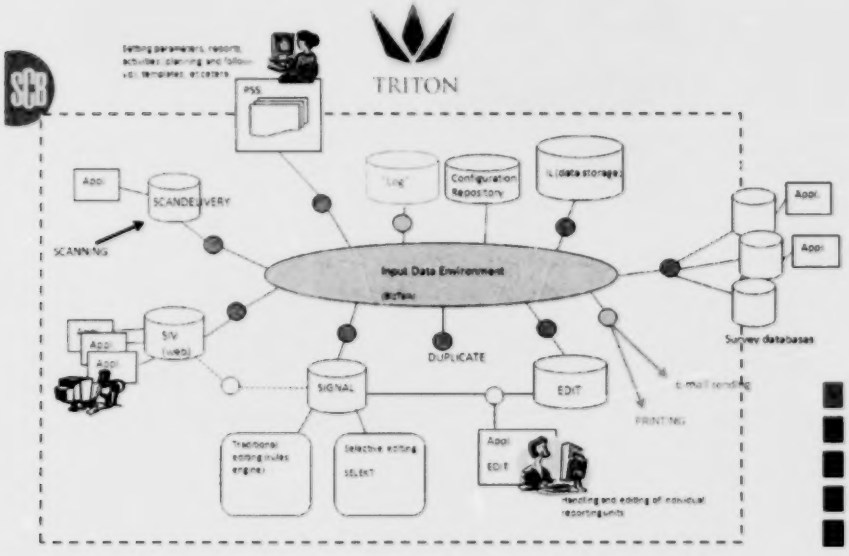
Figure 1
The Swedish adaptation of GSBPM

Figure 2
A common production environment



5. It was decided to start the development at the data collection part of the production process. There were several reasons for this, the main one being that many systems were outdated and that with data collection being a staff-intensive part of the production process, the potential gains in efficiency were large. Also, the integration with our standard tools for web collection, scanning etc. was not very efficient, and there was a need to eliminate manual handling in the production to minimise the risk of errors. Furthermore, we had a new tool for selective editing (SELEKT) which needed to be integrated with an environment for investigation of potential errors. A project was initiated to build a common production environment for data collection that could fulfil these aims, and that could also lay the foundation for a common production environment on a larger scale. This project was TRITON. The first version of TRITON was launched in the middle of 2011, and since then a number of projects and releases with further development have been carried out, also in many ways taking the gradual implementation of surveys in TRITON into account, and re-building parts and infrastructure where necessary. At the moment, Statistics Sweden is already using version 4.8 of TRITON and has around 30 surveys in the system, with another 20 due to be developed before the first quarter of 2016.

Figure 3
TRITON



6. The main parts of TRITON can be shown and listed as follows:

- The Input Data environment that steers the information flows between the different parts
- The data storage (IL) which has been re-created as "shards" or identical databases for each survey and production round to gain independence and performance
- Connections to data coming in through web data collection and scanning – data collection through telephone interviews
- Editing through an error localisation service (SIGNAL) and a connection to SELEKT

- An application (EDIT) for treatment of reporting units – collection, investigation of potential errors, communication etc. This tool has been enhanced greatly since the first version, including several ways to look at data, handling changes in reporting units and more
- Printing and sending of letters and e-mails to reporting units
- Delivery of data to other systems
- And – last but not least – the new version of PSS called "Process areas".

7. Process areas are a kind of dashboard to look at production. They give access to all the tools needed, contain process data reports on the current status and include a quality assurance and control system. The process areas and the overall architecture of TRITON are the tools that can be re-used in other parts of the common production environment. The process areas are built in a hierarchical structure with three levels: Survey, Production round and (for data collection) Collection round. In parallel to Collection rounds we envision two more process areas at the lowest level: Calculation/Processing round and Dissemination round (these names are preliminary and subject to change). Introducing Process areas is also taking a large step towards an active PSS, since the content of a Process area for a specific survey is dependent on the design choices of that survey. This is achieved by letting the survey fill in a questionnaire detailing its main design choices, after which the area is generated with content based on the information in that questionnaire.

8. In 2012, discussions started on where to go next in expanding the common production environment. There were two main options – either to continue with the processes coming after collection and micro editing (further processing and analysis) or to take the next step in the dissemination area where the process is already very functionalised (as is data collection) and where it would probably be easier to move forward. Since the IT tools for dissemination were also undergoing changes, some being modernised and others being replaced, it was decided to start in that area and to begin with a pre-study in the Process and Analysis area, where the way forward was a little more uncertain, and where a recently-started project dedicated to a number of surveys in short-term economic statistics could be used as a first step to see how to proceed.

9. The project in the dissemination area started in 2013 with identification of the main IT tools that would be replaced, and with discussions about how the process area structure would function for dissemination. It was agreed that the structure used for data collection would fit for dissemination as well. However, it was decided not to proceed immediately with introducing process areas for dissemination, but rather to focus on replacing the most important IT tools first. In 2014, the metadata structure for dissemination was put in place in the new tool "PX Meta", and in 2015 work is ongoing with a tool for planning publication on the Statistics Sweden website. Later this year some work on process areas might be undertaken.

10. In the Process and Analyse area, the project for short-term economic statistics has developed a tool for planning and running calculation, imputation, estimation and more. The tool is called KLON, and while not built to be a completely generic tool, it has the potential to become one. A bridge has also been built between TRITON and KLON so that data collected in TRITON is transferred automatically to KLON. But metadata-wise, KLON and TRITON are not fully integrated, meaning some set up of metadata is duplicated. All micro data editing is done in TRITON (with the EDIT tool). The first version of KLON was finished in early 2015. A few more products have already decided to move their calculations into KLON. This has meant that the planned pre-study on how to proceed with the common production environment for Process and Analyse was carried out on time and started late 2014 with the aim of describing how much of KLON could be re-used when expanding to a more generalised production environment. While conclusions

were a bit difficult to draw given that the pre-study was done before KLON version 1 was finished, it can be seen that several things from KLON will be very re-usable. It also became clear that there were somewhat differing opinions on how far it was possible to go with a common production environment for Process and Analyse, and which were the most important things to develop, so further internal discussions need to be held on this matter.

11. At the end of 2014, with all the different initiatives ongoing and with old systems still needing to be replaced, a new approach to the project was taken. An initiative to merge all the different parts of development into one project, also incorporating implementation in a number of surveys and replacing their old systems, has been taken in 2015. This large project is still being set up and organised at the time of writing, and will be fully operational in the second half of 2015, with the objective of leading the way to a common production environment for all steps of the production process from data collection to dissemination in 2015, 2016 and beyond. Beside IT development and implementation, a third part of this project will deal with the overarching structures, including architectural issues, common metadata (that are to be used across processes) and structures for the PSS process areas, among other things. Examples of common metadata are data on surveys, production rounds etc., variables and value sets including classifications, data deliveries and which survey uses which data. This is done to ensure that, in spite of the likelihood that we will opt for three different parts (data collection, processing and analysis, dissemination) within the common production environment, which may differ in some aspects, the three parts will fit together to eliminate duplication of metadata etc. To ensure this, the Generic Statistical Information Model (GSIM) will be used as much as possible. A comparison between GSIM and the information model used in TRITON has been carried out, and has not found large differences, but has rather pinpointed the need for a common understanding of the GSIM objects in the different processes in order to have an integrated environment while allowing each part to be built to satisfy the needs of the specific process. Much important common metadata is found in the "blue" business part of GSIM.

12. The TRITON architecture has also been used in a proof-of-concept project within the UNECE initiative of the Common Statistical Production Architecture (CSPA), successfully receiving answers to interviews from the computer-assisted survey processing, BLAISE, and sending them through the platform. This has strengthened the feeling that the architecture upon which our environment is built is strong and expandable.

III. Reflections: Issues and challenges

13. Building an IT platform for all parts of the statistical production process takes a lot of effort, resources and time. In the meantime, there are many other challenges that need to be met within a statistical office. The path Statistics Sweden has taken is based on taking one process at a time, and also working in three different areas in rather different ways. While the process is moving in the right direction at a steady pace, there have also been challenges.

14. The three main areas in the production process are now differently organised. At Statistics Sweden, staff in charge of data collection and dissemination are to a large extent centralised, whereas processing and analysing is mostly carried out by survey managers and methodologists survey by survey. Collection is highly dependent on IT tools to get an overview of the status of production. Dissemination on the website also uses many IT tools that are common across all surveys, while in process and analyse there are several tools that are useful but which are mostly used in combination with survey specific adaptations. This in turn is partly due to surveys having very different data storage solutions, i.e. there is not a common data warehouse for all surveys. This has meant that there is a need to take these considerations into account when building the common production environment; all parts

should not work in exactly the same way. On the other hand, a survey manager wants to have an overview of the production from start to finish, so the environment needs to incorporate that need as well. The main challenge is to make each process feel part of an integrated whole while supporting the specific needs of different processes.

15. Statistics Sweden has worked by development in stages, both between processes (data collection, then dissemination, then processing and analysing) and within processes (TRITON has been developed in stages and is still not "finished"). This has both pros and cons. The biggest pros are that results are revealed step-by-step rather than having to wait for a "big bang" when everything is finished, and that experiences from early implementations can be taken into account in further development. In fact, we strongly believe that waiting for a complete new system for all areas of production to be finished before releasing it would probably mean it never being finished at all. The biggest con is probably that one has to work with temporary solutions for a long while, combining the new common environment with parts and systems from older environments. Furthermore, the first versions released may be deemed "not good enough" even though further development is already planned and ongoing. Also, there is often a wish from different parts of the organisation to proceed fast in their specific area, making prioritisation difficult.

16. Moving from survey specific systems to a common production environment is a large step, and a common environment necessarily becomes a compromise between different demands from surveys used in very different production systems. There is a need to find solutions that are good enough for most while not moving all possibilities from all old systems into one big system instead, in order to make the new solution easy to maintain and to permit high performance. On the other hand, the new environment needs to be flexible to accommodate different kinds of surveys, not making production unnecessarily difficult for some kinds of surveys. Finding how much functionality to incorporate in an optimal way is a great challenge. The development and the strategy also needs strong support from upper management all across the office, because if asked to choose between having a completely custom-built survey specific system and joining a common environment, many survey managers would choose the first alternative if they had the economic possibility to do so. If the office wants as many surveys as possible to use the common production environment, it must be supported even if every demand from every survey cannot be perfectly met.

17. Implementation of the new environment is a fundamental step that might be easy to overlook. Surveys need to start implementing early to get a feel for the new system and to be ready in time for production. On the other hand, changes and adaptations take place right up to the last minute. Support in the implementation process is of utmost importance, and very different to IT development, so it needs to be led by other people than the developers. Also, post-implementation support is very important when production is running. In both these areas we have struggled a little to meet the needs of the surveys. There is also an in-built conflict between further development (adding new functionality) and maintaining what is already there, since the same staff and resources are needed for both. So we would advise any country that is going on the same journey to consider this very carefully and to devote necessary resources for it. In our case, it is also a journey from the small survey specific systems to large cross-office systems, with the inherent risk of being a "single point of failure". Resources must be devoted to the large systems because so many people are dependent on them working all the time, which is a different situation to when each survey has its own system.

18. Even though we have faced a few challenges, encountered some obstacles and progress has been a bit slower than anticipated when we started, we are still convinced that this is the right way to go. The updated strategy for Statistics Sweden ("Strategy 2020") also clearly points out the common production environment as an important means to

achieve our goals of producing high quality statistics based on the needs of our users, while easing the burden for respondents.